Network AI –
Machine Learning for Irregular Time Series (ML4ITS)
https://ml4its.github.io

Motivation

Time series data is **ubiquitous**. The broad diffusion and adoption of the Internet of Things, along with the major advances in sensor technology, are examples of why such data has become so pervasive. These technologies have applications in several domains, such as **healthcare**, **finance**, **meteorology**, and **transportation**, where related tasks on time series are of high importance. Time series data influences both political and industrial decisions every day, yet there is, surprisingly, very limited research in Machine Learning (ML) for time series - especially in situations where **data is scarce or of low-quality.**

In many real-world applications, we have the following two scenarios:
1. **limited amount of available training data**
2. **huge amount of available data which is scarcely or not labelled due to high costs** of collecting and annotation of the data.

As a result, the future of Artificial Intelligence (AI) will be "about less data, not more". Thus, there is a need for focusing on modern AI techniques that can extract value from small datasets. These considerations can also contribute to the increasing need to address the **sustainability and privacy aspects** of ML and AI.

Goal

The goal of this project is to overcome the issue of limited available or labelled data for (multivariate) time series modelling, where heterogeneity of the data (e.g. non-stationarity, multi-resolution, irregular sampling) as well as noise, pose further challenges.

ML4ITS's main objective is to **advance the state-of-the-art in time series analysis for "irregular" time series**.

We define time series to be “irregular” if they fall under one or several of the following categories:

- Short: univariate/multivariate time series with a limited amount of data and history.
- Multiresolution: multivariate time series where each signal has a different granularity or resolution in terms of sampling frequency.
- Noisy: univariate/multivariate time series with some additional perturbation appearing in different forms. In this class, we also include time series with missing data.
- Heterogeneous: multivariate time series, usually collected by many physical systems, that exhibit different types of embedded, statistical patterns and behaviors.
- Scarcely labelled and unlabeled: univariate/multivariate time series where only a small part of the data is labelled or completely unlabeled.

These are the characteristics of time series data coming from real business cases.

The project aims to **develop methodology** that handles irregular time series for the following **tasks**:

- Forecasting: predicting the future values of the time series based on current/past data;
- Imputation/denoising: creating “clean” data for missing or noisy data cases;
- Anomaly detection and failure prediction: individuating when signals are unusual or indicating that a system is in a critical state;
- Synthetic data creation: among its scopes, addresses the need for creating datasets that are privacy preserving, including the quantification of uncertainties for each task in question. We plan to achieve these goals by developing novel Deep Learning methods in the areas of Transfer Learning, Unsupervised learning and Data Augmentation. These techniques have hardly been explored in the time series domain.

Outcomes and impact

ML4ITS will provide data analysis solutions for time-series collected during real-life industrial operations. Industrial time-series are much more challenging than the thoroughly cleaned and curated time-series used as benchmarks by the ML community, presenting properties such as size (having only short series makes the learning of the dynamics difficult); varying sampling frequencies; high level of noise that do not follow standard statistical distributions; heterogeneity; lack of gold-standard labelling for classification tasks.

ML4ITS will develop new technologies for such irregular time-series, that will significantly advance the state of the art within transfer learning, unsupervised learning, and data augmentation.

The validity of the developed methods will be evaluated through a series of challenging and high-impact use-cases from different domains. The use-cases, covering FinTech, Energy, IoT and Telco, are selected due to their importance for Norwegian competitiveness.

Telenor use cases

1. **Internet of Things**. Availability of low-cost sensors to collect data is instrumenting the physical world, and most of the data coming from these sensors is in the form of (multivariate) time series. Sensor data can be used for automation, recommendation and decision making, which, in turn, will involve time series prediction and anomaly detection through sequence models, as well as providing explainability of the outputs. The use-case in this area will be air quality monitoring and prediction, and the data providers will be public actors including the Norwegian Environment Agency (pollutants), the Norwegian Meteorological Institute (weather), the Norwegian Public Roads Administration (traffic) as well as Trondheim municipality and Telenor (pollutant data from microsensors and population data). This use-case will be run as a collaboration with and as an extension of the AI4IoT pilot in the AI4EU project (2019-2021), where both Telenor and NTNU are partners.

2. **RAN/Core Network anomaly detection and failure prediction.** In the Telco domain, efficient and accurate Anomaly Detection is vital to be able to continuously monitor the base station's key metrics and alert for possible incidents in time. Most commonly, the anomalies to be detected come from systems recording several counters, that is, generating multivariate time series. The difficulty in individuating anomalies in multivariate time series arises from the fact that the contexts and correlations between the different features and time windows have to be taken into account and examined.

   There are two main types of anomalies that are desirable to detect:
   1) point anomalies
   2) trend anomalies.

   The latter, corresponding to failures in the Network, are especially hard to recognize, as they do not differ much from the "normal" situation, so one has to rely on advanced state-of-the-art deep learning models in order to discover them.

3. **Synthetic data generation for overcoming privacy issues in sensitive customer data.** Privacy concerns are one of the main obstacles when processing data to improve the network, customer care and optimize customer interaction. Developing a generative model that produces synthetic data which in terms of its characteristics doesn't differ too much from the original data will solve the privacy problem by still being able to extract value from it.

4. **5G roll out: prediction of traffic and demand.** With the advent of 5G there will correspondingly be a change in the infrastructure and an increase in traffic. To be able to predict the future outcomes that 5G will bring there is a need to encode relevant 4G-related time series information and generate hypothetical scenarios where 4G and 5G networks co-exist. This will be based upon what impact the roll out of 5G will have in small trial areas, and will support the Telco operator's business decisions.

Project structure and Communication

This project relies on a multidisciplinary approach, which is why the ML4ITS consortium has been created with researchers from NTNU, Telenor, FinTech and IoT industries.

The multiple applications of this project primarily concern the areas of IoT, Network automation and financial markets. Several students are contributing to the goals of the project, four PhDs were hired and there are also several Master students. A Hackathon was also organized to get ideas and inspiration for one of Telenor's problems. The winning team proposed a solution on how to detect failures in the network.

Contact person

Sara Malacarne, sara.malacarne@telenor.com